Propuesta de arquitectura de software con IA para monitoreo de cumplimiento en operaciones de tesorería

Documento técnico de solución

Enfoque del documento
La solución propuesta no se plantea como un chatbot aislado. El componente conversacional puede existir como capa de consulta, pero el núcleo debe ser un motor de cumplimiento auditable que relacione operaciones, medios verificables, transcripciones, reglas, metodologías, alertas y evidencias.

Confidencial - Uso interno

Contenido

- 1. Resumen ejecutivo
- 2. Diagnóstico técnico directo
- 3. Objetivos de la solución
- 4. Alcance funcional
- 5. Fuentes de información e integración
- 6. Arquitectura propuesta
- 7. Flujo operativo end-to-end
- 8. Núcleo de inteligencia artificial
- 9. Motor de reglas y metodologías
- 10. Mapeo de metodologías a módulos automatizables
- 11. Modelo de datos lógico
- 12. Seguridad, auditoría y gobierno
- 13. Infraestructura y despliegue
- 14. MLOps, LLMOps y operación
- 15. Casos de uso principales
- 16. API e integración
- 17. Plan de implementación por fases
- 18. Riesgos técnicos y mitigaciones
- 19. Criterios de aceptación
- 20. Diagramas en PlantUML
- 21. Conclusión técnica

1. Resumen ejecutivo

El área de cumplimiento necesita revisar operaciones de compra y venta ejecutadas por traders, validar si se ajustan a reglas regulatorias, revisar medios verificables, comparar horarios de cierre contra horarios registrados y detectar patrones atípicos. El proceso actual es manual, consume demasiado tiempo y no escala frente al volumen de operaciones, llamadas, chats y registros provenientes de diferentes sistemas.

La solución recomendada es una plataforma de monitoreo de cumplimiento asistida por inteligencia artificial. La IA debe utilizarse para transcribir llamadas, extraer entidades relevantes, relacionar comunicaciones con operaciones, priorizar alertas, generar resúmenes explicables y asistir al analista. La decisión final debe permanecer en manos del equipo de cumplimiento, porque el sistema operará en un contexto regulado, sensible y auditable.

El diseño propuesto separa claramente tres capas: integración de datos, motor analítico y experiencia de usuario. La capa de integración captura información desde NICE, Bloomberg, Set-FX, Summit, Master Trader, DataMart, Excel y correos autorizados. La capa analítica normaliza los datos, ejecuta reglas, modelos estadísticos, análisis de lenguaje y búsqueda semántica. La capa de usuario presenta alertas, evidencias, trazabilidad y reportes.

2. Diagnóstico técnico directo

El error de enfoque sería creer que el problema se resuelve con un chatbot. Un chatbot por sí solo no garantiza trazabilidad, evidencia, auditoría, cumplimiento normativo ni control de falsos positivos. En un banco, una respuesta bonita no sirve si no se puede demostrar de dónde salió, qué regla se aplicó, qué evidencia la soporta y quién tomó la decisión final.

La solución correcta debe parecerse más a una fábrica de control y auditoría que a una aplicación conversacional. La IA debe estar encerrada dentro de un flujo controlado: ingesta, normalización, transcripción, extracción, asociación, reglas, alertas, revisión humana y evidencia.

- **Punto crítico 1:** sin datos normalizados, la IA no tendrá una base confiable para analizar operaciones.
- **Punto crítico 2:** sin timestamps precisos en llamadas y chats, no se puede comprobar registro oportuno.
- **Punto crítico 3:** sin trazabilidad de evidencia, el sistema no será defendible ante auditoría interna, revisoría o regulador.
- **Punto crítico 4:** sin un motor de reglas parametrizable, cada cambio normativo obligará a modificar código.
- **Punto crítico 5:** sin revisión humana, el riesgo de decisiones automáticas erróneas es demasiado alto.

Decisión arquitectónica recomendada
Construir una plataforma web de monitoreo de cumplimiento con IA en el núcleo analítico. El chatbot debe ser una interfaz de consulta, no el corazón de la solución.

3. Objetivos de la solución

3.1 Objetivo general

Diseñar una arquitectura de software basada en IA, reglas de negocio y analítica de datos para automatizar el monitoreo de operaciones de tesorería, relacionando operaciones con medios verificables y generando alertas explicables para el equipo de cumplimiento.

3.2 Objetivos específicos

- Transcribir automáticamente llamadas provenientes de NICE y conservar marcas de tiempo por segmento.
- Ingerir registros operativos desde Set-FX, Summit, Master Trader, DataMart, Excel y otras fuentes autorizadas.
- Ingerir chats de Bloomberg u otros medios escritos cuando exista acceso técnico y legal autorizado.

- Normalizar operaciones en un modelo canónico común para comparar datos provenientes de sistemas diferentes.
- Relacionar cada operación con sus comunicaciones asociadas: llamada, chat, correo o medio verificable.
- Ejecutar reglas de cumplimiento según metodología, periodicidad, producto, trader, mesa, contraparte y horario.
- Detectar operaciones atípicas por precio, spread, modificación, anulación, atribución, vínculo o registro extemporáneo.
- Presentar alertas priorizadas con evidencia, explicación, severidad y recomendación de revisión.
- Generar reportes periódicos para cumplimiento, auditoría y gestión interna.
- Permitir que los analistas validen, descarten, escalen o documenten cada alerta.

4. Alcance funcional

4.1 Dentro del alcance

- Captura y procesamiento de llamadas grabadas.
- Transcripción automática de audio a texto.
- Identificación de participantes, palabras clave, montos, monedas, tasas, producto, hora de cierre y contraparte cuando sea posible.
- Integración con fuentes operativas mediante APIs, consultas controladas, archivos planos, exportaciones o DataMart.
- Motor de reglas para metodologías de cumplimiento.
- Modelos estadísticos para detección de operaciones alejadas del precio de referencia.
- Panel web para revisión de alertas, casos y evidencias.
- Búsqueda semántica sobre transcripciones, chats y documentos.
- Generación de informes y bitácora auditable.
- Gestión del ciclo de vida de una alerta: nueva, en revisión, confirmada, descartada, escalada y cerrada.

4.2 Fuera del alcance inicial

- Reemplazar completamente al analista de cumplimiento.
- Tomar decisiones sancionatorias automáticas.
- Modificar directamente operaciones en sistemas core.
- Acceder a comunicaciones sin autorización formal.
- Entrenar modelos con datos sensibles sin controles de privacidad, retención y gobierno.
- Integración profunda con vendors sin acuerdos técnicos, contractuales y de seguridad.

4.3 Principio de diseño

El sistema debe operar como un asistente de control. Debe encontrar, ordenar y explicar señales de riesgo; no debe actuar como juez definitivo. En cumplimiento financiero, la automatización debe aumentar cobertura y trazabilidad, no eliminar responsabilidad humana.

5. Fuentes de información e integración

Fuente	Datos relevantes	Mecanismo de integración	Uso en cumplimiento
NICE	Llamadas grabadas, metadatos de audio, duración, fecha, participantes si están disponibles.	Conector a repositorio de grabaciones, exportación controlada o API del vendor.	Base para transcripción y medio verificable.
Bloomberg	Chats, confirmaciones, conversaciones de cierre.	API, exportación certificada o integración con proveedor.	Medio verificable escrito y evidencia conversacional.
Set-FX	Operaciones de divisas, hora de registro, producto, moneda,	API, consulta autorizada,	Validación de registro oportuno y comparación

	monto y contraparte.	DataMart o archivo de cierre.	contra cierre real.
Summit	Operaciones, versiones, anulaciones, modificaciones, aprobaciones y atribuciones.	DataMart, extracción diaria o conexión controlada.	Control de anuladas/modificadas, atribuciones y trazabilidad.
Master Trader	Registros operativos y horarios.	API, base de datos, archivos o reportes.	Cruce con medios verificables.
DataMart	Consolidado operativo usado por cumplimiento.	Consultas programadas, vistas seguras o jobs ETL.	Fuente inicial recomendada para MVP por menor fricción.
Excel / macros	Muestras, reportes manuales, parámetros y resultados históricos.	Carga controlada, validación de esquema y versionamiento.	Migración gradual de controles manuales.
Correo corporativo	Ratificaciones, aprobaciones, autorizaciones y soporte de atribuciones.	Conector autorizado con criterios de búsqueda.	Evidencia documental y aprobación excepcional.

5.1 Estrategia de integración por madurez

- **Fase inicial:** usar DataMart, exportaciones y archivos controlados para probar valor rápidamente.
- **Fase intermedia:** crear conectores estables hacia sistemas internos y repositorios de medios verificables.
- **Fase avanzada:** integrar APIs de vendors con gobierno, monitoreo, contratos de datos y trazabilidad completa.

6. Arquitectura propuesta

La arquitectura recomendada es modular, orientada a eventos y con separación entre datos crudos, datos normalizados, reglas de cumplimiento, modelos de IA y experiencia de usuario. Esta separación permite cambiar una metodología sin romper todo el sistema, auditar cada resultado y escalar procesamiento de audio o reglas de forma independiente.

6.1 Capas de arquitectura

Capa	Componentes	Responsabilidad
Capa de captura	Conectores hacia NICE, Bloomberg, Set-FX, Summit, Master Trader, DataMart, Excel y correo.	Traer datos y medios verificables sin alterar los sistemas fuente.
Capa de datos crudos	Data Lake / almacenamiento raw.	Guardar archivos originales, audios, chats y exportaciones sin pérdida.
Capa de normalización	ETL/ELT, validación de esquema, limpieza, deduplicación y modelo canónico.	Convertir datos heterogéneos en estructuras comparables.
Capa de IA	Transcripción, extracción NLP, clasificación, búsqueda semántica, resúmenes y ranking de riesgo.	Convertir comunicaciones no estructuradas en señales útiles.
Capa de reglas	Motor de metodologías, parametrización normativa y validaciones determinísticas.	Aplicar controles de cumplimiento de forma explicable.
Capa analítica	Detección estadística de atípicos, percentiles, tendencias, scoring y priorización.	Detectar patrones difíciles de cubrir manualmente.
Capa de casos	Alertas, workflow, asignación, comentarios, evidencias y estados.	Gestionar la revisión humana y documentar decisiones.

Capa de presentación	Dashboard web, reportes, consultas y chatbot asistente.	Entregar información accionable a cumplimiento.
Capa transversal	Seguridad, auditoría, observabilidad, gobierno, permisos y retención.	Proteger datos sensibles y soportar auditoría.

6.2 Componentes principales

- **Orquestador de ingesta:** programa cargas diarias, semanales y trimestrales según cada metodología.
- **Servicio de transcripción:** procesa audios, separa segmentos, conserva timestamps y genera texto auditable.
- **Servicio de extracción NLP:** detecta entidades: producto, moneda, tasa, monto, contraparte, intención de cierre, hora y palabras clave.
- **Motor de asociación de evidencias:** vincula operación con llamada, chat o correo según trader, fecha, hora, contraparte, producto, monto y similitud semántica.
- **Motor de reglas:** ejecuta validaciones determinísticas como registro dentro de 15 minutos, límites de atribución o existencia de aprobación.
- **Motor estadístico:** calcula percentiles, spreads, desviaciones, niveles de atípico y tendencias.
- **Gestor de alertas:** crea casos, asigna severidad, prioriza y controla estados.
- **Repositorio de evidencias:** almacena transcripción, audio original, chat, operación, versión y explicación del hallazgo.
- **Panel de cumplimiento:** permite revisar, filtrar, escuchar, leer, comparar y cerrar alertas.
- **Asistente conversacional:** responde preguntas sobre casos, evidencia y metodología, siempre citando datos internos y sin inventar conclusiones.

7. Flujo operativo end-to-end

1. El sistema captura operaciones desde DataMart, Set-FX, Summit, Master Trader o archivos autorizados.
2. Se capturan medios verificables: llamadas NICE, chats Bloomberg, correos o documentos de soporte.
3. Los datos crudos se guardan sin alteración en una zona raw para trazabilidad.
4. Los audios se transcriben con marcas de tiempo y, si es viable, diarización por hablante.
5. Las transcripciones y chats se procesan para extraer entidades: producto, monto, moneda, tasa, hora, contraparte, intención de cierre y expresiones de confirmación.
6. Las operaciones se transforman a un modelo canónico común.
7. El motor de asociación busca coincidencias entre operación y comunicación.
8. El motor de reglas ejecuta las metodologías aplicables según producto, periodicidad, trader, mesa y tipo de operación.
9. El motor estadístico detecta operaciones alejadas del mercado o patrones inusuales.
10. El sistema genera alertas con explicación, evidencia y nivel de prioridad.
11. El analista revisa la alerta en el panel, escucha segmentos relevantes, lee transcripción y valida la evidencia.
12. El analista confirma, descarta, escala o cierra el caso.
13. El sistema conserva bitácora, comentarios, evidencia y resultado para auditoría.

8. Núcleo de inteligencia artificial

La IA debe utilizarse donde realmente aporta valor: transformar comunicaciones no estructuradas en datos revisables, priorizar señales y acelerar la investigación. No debe usarse para reemplazar reglas regulatorias determinísticas ni para inventar interpretaciones sin evidencia.

Módulo IA	Función	Técnica posible	Salida
Transcripción automática	Audio a texto con timestamps.	OpenAI Whisper u otro motor STT validado internamente.	Transcripción, segmentos, confianza, archivo fuente.
Diarización	Separación aproximada de	Modelo de diarización o	Hablante 1, hablante 2, posibles

	hablantes.	metadatos de llamada.	roles.
Extracción de entidades	Identificar monto, moneda, tasa, producto, fecha, hora, contraparte y verbos de cierre.	NER, reglas lingüísticas, expresiones regulares, LLM controlado.	Entidades estructuradas con evidencia textual.
Clasificación de intención	Detectar confirmación, negociación, modificación, anulación o ratificación.	Clasificador supervisado o LLM con prompt restringido.	Tipo de evento comunicacional.
Búsqueda semántica	Encontrar llamadas/chats relacionados con una operación aunque no coincidan palabras exactas.	Embeddings y base vectorial interna.	Ranking de evidencias candidatas.
Resumen explicable	Crear resumen del caso para el analista.	LLM con RAG sobre evidencia interna.	Resumen con citas internas y advertencias de incertidumbre.
Scoring de riesgo	Priorizar alertas combinando reglas, atípicos y evidencia incompleta.	Modelo interpretable o score ponderado.	Severidad y razones.

8.1 Reglas de uso seguro de IA

- Todo resultado de IA debe guardar evidencia de origen.
- El modelo debe informar incertidumbre cuando no pueda asociar comunicación y operación.
- Los resúmenes no pueden ser la única fuente de decisión.
- Las reglas regulatorias no deben depender exclusivamente de un LLM.
- Los prompts, versiones de modelo y parámetros deben versionarse.
- Debe existir revisión humana obligatoria para alertas críticas.
- Los datos sensibles no deben enviarse a servicios externos sin aprobación legal, contractual y de seguridad.

9. Motor de reglas y metodologías

El motor de reglas es el centro determinístico de la solución. Su función es ejecutar las metodologías de cumplimiento con parámetros claros, versionados y auditables. La IA alimenta al motor con datos extraídos, pero la evaluación de cumplimiento debe quedar expresada en reglas entendibles.

9.1 Diseño recomendado del motor de reglas

- **Reglas parametrizables:** por ejemplo, ventana máxima de registro de 15 minutos, productos aplicables, excepciones, periodicidad y criterios de muestreo.
- **Versionamiento:** cada regla debe tener versión, vigencia, autoría funcional y fecha de activación.
- **Explicabilidad:** cada alerta debe indicar qué regla falló, con qué datos y contra qué evidencia.
- **Separación técnica:** las reglas deben vivir fuera del código duro, idealmente en tablas, YAML o una interfaz administrativa.
- **Trazabilidad:** cada ejecución debe guardar input, output, versión de regla y timestamp.

9.2 Ejemplo de estructura de regla

```
regla:
  id: RO-001
  nombre: Registro oportuno en Set-FX
  version: 1.0
  vigencia_desde: 2026-01-01
  periodicidad: diaria
  aplica_a:
    productos: [Divisas, Derivados]
    sistemas: [Set-FX]
  condicion:
    diferencia_minutos: hora_registro - hora_cierre_medio_verificable
    maximo_permitido: 15
  excepciones:
    - moneda_no_incluida_en_base
    - monto_menor_a_250000_usd_equivalente
  resultado:
```

```
  si_cumple: Sin alerta
  si_no_cumple: Alerta registro extemporaneo
evidencia_requerida:
  - operacion
  - medio_verificable
  - transcripcion_o_chat
  - calculo_diferencia_minutos
```

10. Mapeo de metodologías a módulos automatizables

Metodología	Periodicidad	Automatización propuesta	Fuentes	Salida
Aleatorio Presencial	Trimestral	Selección aleatoria de día, extracción de comunicaciones y revisión de todos los cierres del trader.	Operaciones, llamadas, chats, calendario/presencialidad si existe.	Muestra trimestral, trazabilidad de cierre y comunicaciones.
Aleatorio Registro oportuno	Semanal / diario en ejecución	Selección aleatoria por trader y validación de hora de registro contra hora de cierre en medio verificable.	Set-FX, Master Trader, llamadas, chats.	Alerta si excede ventana permitida o si no hay evidencia suficiente.
Aleatorio Trabajo Remoto	Trimestral	Selección de dos días remotos no consecutivos y representativos según promedio de operaciones.	Operaciones OTC, calendario de trabajo remoto, llamadas, chats.	Casos de revisión por día/trader con evidencia completa.
Análisis de posturas	Diario	Detección de señales en posturas ingresadas y cruce con medios verificables.	Sistemas de mercado, posturas, operaciones, comunicaciones.	Alertas por posturas sospechosas o inconsistentes.
Anuladas y modificadas	Diario	Comparación de versiones finales T-1 contra T-2 por producto y seguimiento de tendencias.	Summit, DataMart, versiones de cierre.	Alertas por modificación, anulación, frecuencia o patrón inusual.
Atribuciones por operador	Diario	Validación de operaciones contra límites, certificaciones, autorizaciones y aprobaciones.	Summit, SARM, AMV, correo, maestro de operadores.	Alertas por operación fuera de atribución o aprobación insuficiente.
Registro Oportuno	Semanal con análisis diario	Consolidado de operaciones extemporáneas, revisión de medio verificable y motivo del retraso.	Set-FX, llamadas, chats, excepciones, DataMart.	Listado consolidado de incumplimientos y justificaciones.
Revisión de tasas de negociación	Diario	Cálculo de spreads, percentiles, desviaciones y clasificación de atípicos por filtros 3A, 3B, 3C, 3D y 5.	Operaciones, tasas de mercado, límites, devaluación, libros y segmentos.	Alertas leve/extremo/no alerta con cálculo explicable.
Vinculados y partes relacionadas	Semanal	Cruce de contrapartes contra listas de vinculados y políticas aplicables.	Operaciones, maestro de contrapartes, listas internas.	Alertas por contraparte vinculada o política incumplida.

10.1 Priorización por valor de implementación

Prioridad	Control	Valor	Complejidad	Motivo
1	Registro oportuno	Alto	Alta	Tiene regla clara: comparar hora de cierre contra hora de registro.
2	Anuladas y modificadas	Alto	Media	Requiere comparación T-1/T-2 y tendencias, pero depende más de datos que de IA.
3	Revisión de tasas de negociación	Alto	Media/Alta	Aporta detección de riesgo, pero necesita datos de mercado y calibración estadística.
4	Atribuciones por operador	Alto	Media	Necesita maestro confiable

				de límites, certificaciones y autorizaciones.
5	Aleatorios presencial/remoto	Medio	Media	Automatiza muestreo y recolección de evidencia, reduce carga manual.
6	Vinculados y partes relacionadas	Medio/Alto	Baja/Media	Depende de calidad de listas y políticas internas.
7	Análisis de posturas	Medio	Alta	Requiere definición fuerte de señales y acceso a posturas.

11. Modelo de datos lógico

El modelo de datos debe permitir reconstruir cualquier alerta desde la operación original hasta el resultado del análisis. Por eso se recomienda un modelo basado en operaciones, comunicaciones, evidencias, reglas, alertas y casos.

Entidad	Campos principales
Operacion	id_operacion, sistema_origen, producto, trader, contraparte, moneda, monto, tasa, fecha_cierre, hora_registro, estado, version
Comunicacion	id_comunicacion, tipo, fuente, fecha, hora_inicio, hora_fin, participantes, archivo_original
Transcripcion	id_transcripcion, id_comunicacion, texto, segmento_inicio, segmento_fin, hablante, confianza
EntidadExtraida	id_entidad, tipo, valor, fuente_textual, confianza, timestamp_segmento
Evidencia	id_evidencia, id_operacion, id_comunicacion, tipo_evidencia, score_asociacion, ruta, hash
ReglaCumplimiento	id_regla, nombre, metodologia, version, vigencia, parametros, estado
ResultadoRegla	id_resultado, id_regla, id_operacion, resultado, explicacion, datos_evaluados
Alerta	id_alerta, severidad, tipo, estado, fecha_creacion, score_riesgo, justificacion
Caso	id_caso, id_alerta, analista_asignado, decision, comentarios, fecha_cierre
Auditoria	id_auditoria, usuario, accion, entidad, fecha, antes, despues

12. Seguridad, auditoría y gobierno

Este sistema tratará información sensible: operaciones financieras, comunicaciones grabadas, chats, posibles hallazgos disciplinarios y datos de contrapartes. La arquitectura debe asumir desde el inicio controles fuertes de seguridad y gobierno.

12.1 Controles mínimos

- Autenticación corporativa con SSO y MFA.
- Autorización por roles: administrador, analista, supervisor, auditor, solo lectura.
- Cifrado en tránsito y en reposo.
- Hash de archivos originales para demostrar integridad de evidencia.
- Bitácora inmutable de acciones relevantes.
- Políticas de retención por tipo de dato.
- Enmascaramiento de datos sensibles en ambientes no productivos.
- Control de acceso a audios y transcripciones por necesidad de conocimiento.

- Separación de funciones entre configuración de reglas y revisión de alertas.
- Registro de versión de modelo, versión de regla y datos usados en cada resultado.

12.2 Riesgos de privacidad

- **Audio y llamadas:** pueden contener información personal o confidencial no relacionada con la operación.
- **Chats:** pueden incluir conversaciones sensibles o datos de clientes.
- **LLM externo:** si se usa un servicio externo, debe existir análisis legal, contractual y de seguridad.
- **Datos de entrenamiento:** no se debe entrenar ni ajustar modelos con datos reales sin anonimización, aprobación y controles.

13. Infraestructura y despliegue

La infraestructura puede desplegarse on-premise, en nube privada o en nube pública con controles bancarios. La decisión depende de restricciones de seguridad, residencia de datos, integración con sistemas internos y política del banco.

Opción	Ventajas	Riesgos	Uso recomendado
On-premise	Mayor control de datos y cercanía a sistemas internos.	Mayor complejidad operativa, escalamiento limitado.	Si las llamadas y datos no pueden salir del entorno interno.
Nube privada	Balance entre control, escalabilidad y gobierno.	Requiere madurez de plataforma interna.	Opción recomendada si el banco ya tiene plataforma privada.
Nube pública regulada	Escalabilidad, servicios administrados, IA gestionada.	Requiere revisión legal, seguridad y cumplimiento.	Viable si hay acuerdos de residencia, cifrado y gobierno.
Híbrida	Procesamiento sensible interno y analítica escalable en nube.	Más integración y operación.	Recomendada cuando audio/datos sensibles deben quedarse internos.

13.1 Componentes de infraestructura

- Balanceador o API Gateway.
- Servicios backend en contenedores.
- Cola de eventos para procesamiento asíncrono.
- Workers de transcripción y análisis NLP.
- Base relacional para operaciones, alertas, reglas y casos.
- Almacenamiento de objetos para audios, chats, archivos y evidencias.
- Base vectorial para búsqueda semántica.
- Data warehouse o lakehouse para analítica histórica.
- Servicio de monitoreo, logs y trazas.
- Gestor de secretos y llaves de cifrado.

14. MLOps, LLMOps y operación

La operación de modelos de IA en cumplimiento exige control de versiones, medición de calidad, pruebas, monitoreo de errores y validación humana. Sin esto, el sistema puede parecer avanzado, pero será frágil y difícil de defender.

- **Dataset de validación:** con llamadas, chats y operaciones históricas etiquetadas por cumplimiento.
- **Métricas de transcripción:** tasa de error de palabras, confianza por segmento, calidad por ruido y duración.
- **Métricas de extracción:** precisión y cobertura para monto, tasa, moneda, contraparte, producto y hora de cierre.
- **Métricas de asociación:** porcentaje de operaciones con medio verificable encontrado y score promedio.
- **Métricas de reglas:** alertas generadas, falsos positivos, falsos negativos, tiempo de revisión y severidad.

- **Monitoreo de drift:** cambios en lenguaje, productos, tasas, traders, volumen o calidad de audio.
- **Evaluación de LLM:** pruebas anti-alucinación, consistencia, citas de evidencia y rechazo cuando no hay datos suficientes.

15. Casos de uso principales

ID	Caso de uso	Descripción
CU-01	Procesar llamadas diarias	El sistema captura audios, transcribe, segmenta y guarda evidencia.
CU-02	Validar registro oportuno	Compara hora de cierre detectada en medio verificable contra hora de registro en sistema.
CU-03	Detectar operación atípica por tasa	Calcula spread, percentiles y desviación frente a referencia de mercado.
CU-04	Revisar anuladas y modificadas	Compara versiones de Summit T-1 y T-2, identifica cambios y genera alerta.
CU-05	Validar atribuciones	Cruza operación contra límites, certificaciones y aprobaciones.
CU-06	Buscar evidencia de una operación	Permite consultar llamadas, chats y correos asociados a una operación.
CU-07	Gestionar alerta	Analista revisa, comenta, adjunta evidencia y decide estado del caso.
CU-08	Generar reporte	Consolida resultados por metodología, trader, producto, severidad y periodo.

16. API e integración

16.1 APIs internas sugeridas

Endpoint	Responsabilidad
POST /ingesta/operaciones	Recibe lote de operaciones desde DataMart o sistemas fuente.
POST /ingesta/comunicaciones	Registra audios, chats o correos como medios verificables.
POST /procesamiento/transcribir	Solicita transcripción de un audio.
POST /analisis/asociar-evidencia	Relaciona una operación con comunicaciones candidatas.
POST /reglas/evaluar	Ejecuta reglas de cumplimiento sobre una operación o lote.
GET /alertas	Consulta alertas por estado, severidad, trader, fecha o metodología.
GET /alertas/{id}/evidencias	Obtiene evidencias asociadas a una alerta.
PATCH /casos/{id}	Actualiza estado, decisión, comentarios o asignación.
GET /reportes/metodologias	Genera reporte consolidado por metodología y periodo.

16.2 Contrato lógico de operación normalizada

```
{
  "id_operacion": "string",
  "sistema_origen": "Set-FX | Summit | MasterTrader | DataMart",
  "producto": "Spot | Forward | Swap | Opciones | RentaFija | Liquidez",
  "tipo_operacion": "Compra | Venta | Anulacion | Modificacion",
  "trader": "codigo_interno",
```

```
  "mesa": "Distribucion | Producto | Otra",
  "contraparte": "codigo_contraparte",
  "moneda": "COP | USD | EUR | Otra",
  "monto": 0.0,
  "tasa": 0.0,
  "fecha_negociacion": "YYYY-MM-DD",
  "hora_cierre_reportada": "HH:MM:SS",
  "hora_registro_sistema": "HH:MM:SS",
  "estado": "Registrada | Anulada | Modificada",
  "version": 1,
  "metadatos": {}
}
```

17. Plan de implementación por fases

Fase	Duración estimada	Alcance	Resultado
Fase 0 - Descubrimiento técnico	2 a 4 semanas	Inventario de fuentes, campos, accesos, restricciones legales, muestras de audio y reglas priorizadas.	Mapa de datos, backlog y arquitectura validada.
Fase 1 - MVP Registro oportuno	6 a 10 semanas	Ingesta DataMart/Set-FX, transcripción NICE, asociación básica y regla de 15 minutos.	Primer control automatizado con evidencia.
Fase 2 - Alertas y casos	6 a 8 semanas	Panel web, workflow, bitácora, gestión de evidencias y reportes.	Operación controlada por analistas.
Fase 3 - Anuladas/modificadas y atribuciones	8 a 12 semanas	Integración Summit, comparación de versiones, límites y aprobaciones.	Controles adicionales automatizados.
Fase 4 - Tasas y atípicos	8 a 12 semanas	Modelos estadísticos, percentiles, spreads y clasificación de atípicos.	Motor analítico avanzado.
Fase 5 - Escalamiento	Continuo	Bloomberg, correo, búsqueda semántica, chatbot asistente, monitoreo y hardening.	Plataforma integral de compliance.

18. Riesgos técnicos y mitigaciones

Riesgo	Impacto	Mitigación
Calidad deficiente de audio	Transcripciones incorrectas y falsos hallazgos.	Medir calidad, usar modelos robustos, conservar audio y permitir escucha del segmento.
Datos inconsistentes entre sistemas	No se puede asociar operación con evidencia.	Modelo canónico, reglas de reconciliación y score de asociación.
Falsos positivos altos	El equipo pierde confianza en el sistema.	Priorización, calibración por metodología y revisión de reglas.
Alucinación de LLM	Conclusiones no soportadas por evidencia.	RAG estricto, citas internas, bloqueo cuando no hay evidencia y revisión humana.
Integración con vendors lenta	Retraso del proyecto.	Empezar con DataMart/exportaciones y diseñar conectores reemplazables.
Cambios regulatorios	Reglas quedan obsoletas.	Motor de reglas versionado y parametrizable.
Acceso indebido a datos sensibles	Riesgo legal y reputacional.	RBAC, cifrado, bitácora, segregación y aprobación formal.
Escalamiento de procesamiento	Demoras en transcripción y análisis.	Colas, workers escalables y procesamiento batch/asíncrono.

19. Criterios de aceptación

- El sistema puede cargar operaciones desde al menos una fuente estructurada inicial.
- El sistema puede procesar audios de NICE y generar transcripciones con timestamps.
- Cada alerta generada tiene regla, metodología, datos evaluados y evidencia asociada.
- El analista puede revisar una alerta, consultar la evidencia y dejar decisión documentada.
- El sistema conserva bitácora de acciones y cambios.
- El motor de reglas permite modificar parámetros sin alterar código fuente.
- El sistema permite exportar reportes por fecha, trader, producto, metodología y severidad.
- El porcentaje de operaciones con evidencia candidata asociada mejora frente al proceso manual.
- El tiempo promedio de revisión por operación disminuye de forma medible.
- Los falsos positivos se miden y se reducen mediante calibración.

20. Diagramas en PlantUML

A continuación se incluye código PlantUML para generar los diagramas principales de la solución.

20.1 Diagrama de componentes

```
@startuml
title Arquitectura de componentes - Plataforma IA de Compliance

skinparam componentStyle rectangle
skinparam shadowing false

actor "Analista de Compliance" as Analista
actor "Supervisor" as Supervisor
actor "Auditor" as Auditor

package "Fuentes externas / internas" {
  [NICE\nGrabaciones] as NICE
  [Bloomberg\nChats] as Bloomberg
  [Set-FX] as SetFX
  [Summit] as Summit
  [Master Trader] as MT
  [DataMart] as DM
  [Excel / Macros] as Excel
  [Correo corporativo] as Mail
}

package "Capa de ingesta" {
  [Conectores de datos] as Connectors
  [Validador de esquema] as Validator
  [Orquestador batch/eventos] as Orchestrator
}

package "Capa de datos" {
  database "Data Lake Raw" as Raw
  database "Base operacional normalizada" as ODS
  database "Repositorio de evidencias" as Evidence
  database "Base vectorial" as VectorDB
}

package "Capa de IA y analítica" {
  [Servicio de transcripción] as STT
  [Extracción NLP] as NLP
  [Asociador de evidencias] as Matcher
  [Motor estadístico] as Stats
  [Motor de reglas] as Rules
  [Scoring de riesgo] as Risk
}

package "Capa de negocio" {
  [Gestor de alertas] as Alerts
  [Gestor de casos] as Cases
  [Reportes] as Reports
  [Auditoría] as Audit
}

package "Capa de presentación" {
  [Dashboard web] as Web
  [Asistente conversacional] as Chat
}

NICE --> Connectors
Bloomberg --> Connectors
SetFX --> Connectors
Summit --> Connectors
```

```
MT --> Connectors
DM --> Connectors
Excel --> Connectors
Mail --> Connectors

Connectors --> Validator
Validator --> Orchestrator
Orchestrator --> Raw
Orchestrator --> ODS

Raw --> STT
STT --> NLP
NLP --> VectorDB
NLP --> Matcher
ODS --> Matcher
Matcher --> Evidence

ODS --> Rules
Evidence --> Rules
ODS --> Stats
Stats --> Risk
Rules --> Risk

Risk --> Alerts
Alerts --> Cases
Cases --> Reports
Cases --> Audit

Analista --> Web
Supervisor --> Web
Auditor --> Reports
Web --> Cases
Web --> Evidence
Chat --> VectorDB
Chat --> Cases
Chat --> Reports

@enduml
```

20.2 Diagrama de flujo de datos

```
@startuml
title Flujo de datos - Operaciones, comunicaciones y alertas

start

:Capturar operaciones desde DataMart, Set-FX, Summit o Master Trader;
:Capturar medios verificables desde NICE, Bloomberg o correo;
:Guardar datos originales en zona Raw;
:Validar estructura, duplicados y campos obligatorios;
:Normalizar operación al modelo canónico;

if (Es audio?) then (Sí)
  :Transcribir audio;
  :Segmentar por tiempo y hablante;
else (No)
  :Procesar texto/chat/correo;
endif

:Extraer entidades relevantes;
:Buscar comunicación asociada a la operación;
:Calcular score de asociación;

if (Evidencia suficiente?) then (Sí)
  :Ejecutar reglas de metodología;
  :Ejecutar análisis estadístico si aplica;
  :Calcular score de riesgo;
else (No)
  :Crear alerta por evidencia insuficiente;
endif

if (Existe incumplimiento o señal?) then (Sí)
  :Crear alerta;
  :Adjuntar evidencia y explicación;
  :Asignar a analista;
else (No)
  :Guardar resultado sin alerta;
endif

:Registrar auditoría de ejecución;
stop

@enduml
```

20.3 Diagrama de secuencia - Registro oportuno

```
@startuml
title Secuencia - Validación de registro oportuno

actor "Analista" as A
participant "Dashboard Web" as Web
participant "Gestor de Alertas" as Alertas
participant "Motor de Reglas" as Reglas
participant "Asociador de Evidencias" as Matcher
participant "Servicio NLP" as NLP
participant "Servicio STT" as STT
database "Operaciones Normalizadas" as ODS
database "Repositorio Evidencias" as Evid

A -> Web: Solicita revisión de registro oportuno
Web -> Alertas: Consultar alertas pendientes
Alertas -> Reglas: Ejecutar regla RO-001
Reglas -> ODS: Obtener operación y hora de registro
Reglas -> Matcher: Buscar medio verificable asociado
Matcher -> Evid: Consultar llamadas/chats candidatos
Matcher -> STT: Solicitar transcripción si falta
STT -> NLP: Enviar transcripción segmentada
NLP -> Matcher: Entidades extraídas y hora de cierre
Matcher -> Reglas: Evidencia asociada + hora cierre
Reglas -> Reglas: Calcular diferencia en minutos

alt diferencia <= 15 o existe excepción válida
  Reglas -> Alertas: Resultado sin alerta
else diferencia > 15
  Reglas -> Alertas: Crear alerta registro extemporáneo
  Alertas -> Web: Mostrar alerta, evidencia y explicación
  A -> Web: Revisa y decide
end

@enduml
```

20.4 Diagrama de clases - Modelo de dominio

```
@startuml
title Modelo de dominio - Compliance de operaciones

class Operacion {
  +idOperacion: String
  +sistemaOrigen: String
  +producto: String
  +tipoOperacion: String
  +trader: String
  +contraparte: String
  +moneda: String
  +monto: Decimal
  +tasa: Decimal
  +fechaNegociacion: Date
  +horaRegistro: Time
  +estado: String
  +version: Integer
}

class Comunicacion {
  +idComunicacion: String
  +tipo: String
  +fuente: String
  +fecha: Date
  +horaInicio: Time
  +horaFin: Time
  +archivoOriginal: String
  +hashArchivo: String
}

class Transcripcion {
  +idTranscripcion: String
  +texto: Text
  +segmentoInicio: Time
  +segmentoFin: Time
  +hablante: String
  +confianza: Decimal
}

class EntidadExtraida {
  +idEntidad: String
  +tipo: String
  +valor: String
  +confianza: Decimal
  +fragmento: Text
```

```
}

class Evidencia {
  +idEvidencia: String
  +tipo: String
  +scoreAsociacion: Decimal
  +ruta: String
  +hash: String
}

class ReglaCumplimiento {
  +idRegla: String
  +nombre: String
  +metodologia: String
  +version: String
  +vigenciaDesde: Date
  +parametros: JSON
  +estado: String
}

class ResultadoRegla {
  +idResultado: String
  +resultado: String
  +explicacion: Text
  +datosEvaluados: JSON
  +fechaEjecucion: DateTime
}

class Alerta {
  +idAlerta: String
  +tipo: String
  +severidad: String
  +estado: String
  +scoreRiesgo: Decimal
  +fechaCreacion: DateTime
}

class Caso {
  +idCaso: String
  +estado: String
  +decision: String
  +comentarios: Text
  +fechaCierre: DateTime
}

Operacion "1" -- "0..*" Evidencia
Comunicacion "1" -- "0..*" Transcripcion
Transcripcion "1" -- "0..*" EntidadExtraida
Comunicacion "1" -- "0..*" Evidencia
Operacion "1" -- "0..*" ResultadoRegla
ReglaCumplimiento "1" -- "0..*" ResultadoRegla
ResultadoRegla "0..*" -- "0..1" Alerta
Alerta "1" -- "0..1" Caso
Caso "1" -- "0..*" Evidencia

@enduml
```

20.5 Diagrama de despliegue

```
@startuml
title Despliegue lógico

node "Red corporativa" {
  node "Usuarios internos" {
    [Navegador Analista]
  }

  node "Zona de aplicación" {
    [API Gateway]
    [Backend Compliance]
    [Dashboard Web]
    [Servicio de Reportes]
    [Asistente Conversacional]
  }

  node "Zona de procesamiento" {
    [Workers de Ingesta]
    [Workers STT]
    [Workers NLP]
    [Workers Reglas]
    [Workers Estadísticos]
    queue "Cola de Eventos" as Queue
  }
```

```
  node "Zona de datos" {
    database "Base Relacional" as DB
    database "Data Lake / Objetos" as Lake
    database "Base Vectorial" as VDB
    database "Warehouse Analítico" as WH
    [Gestor de Secretos]
  }

  node "Sistemas fuente" {
    [NICE]
    [Bloomberg]
    [Set-FX]
    [Summit]
    [Master Trader]
    [DataMart]
  }
}

[Navegador Analista] --> [Dashboard Web]
[Dashboard Web] --> [API Gateway]
[API Gateway] --> [Backend Compliance]
[Backend Compliance] --> Queue
Queue --> [Workers de Ingesta]
Queue --> [Workers STT]
Queue --> [Workers NLP]
Queue --> [Workers Reglas]
Queue --> [Workers Estadísticos]

[Workers de Ingesta] --> [NICE]
[Workers de Ingesta] --> [Bloomberg]
[Workers de Ingesta] --> [Set-FX]
[Workers de Ingesta] --> [Summit]
[Workers de Ingesta] --> [Master Trader]
[Workers de Ingesta] --> [DataMart]

[Workers de Ingesta] --> Lake
[Workers de Ingesta] --> DB
[Workers STT] --> Lake
[Workers NLP] --> VDB
[Workers Reglas] --> DB
[Workers Estadísticos] --> WH
[Backend Compliance] --> DB
[Backend Compliance] --> Lake
[Asistente Conversacional] --> VDB

@enduml
```

20.6 Diagrama de estados - Ciclo de vida de alerta

```
@startuml
title Ciclo de vida de una alerta

[*] --> Nueva

Nueva --> Asignada : asignar analista
Asignada --> EnRevision : iniciar revisión
EnRevision --> RequiereEvidencia : evidencia insuficiente
RequiereEvidencia --> EnRevision : evidencia agregada

EnRevision --> Confirmada : incumplimiento validado
EnRevision --> Descartada : falso positivo
EnRevision --> Escalada : requiere decisión superior

Confirmada --> Cerrada : documentar acción
Descartada --> Cerrada : justificar descarte
Escalada --> Cerrada : decisión final registrada

Cerrada --> [*]

@enduml
```

20.7 Diagrama de casos de uso

```
@startuml
title Casos de uso - Plataforma de Compliance

left to right direction

actor "Analista de Compliance" as Analista
actor "Supervisor" as Supervisor
actor "Auditor" as Auditor
actor "Administrador Funcional" as Admin
```

```
actor "Sistema Fuente" as Fuente

rectangle "Plataforma IA Compliance" {
  usecase "Cargar operaciones" as UC1
  usecase "Procesar medios verificables" as UC2
  usecase "Transcribir llamadas" as UC3
  usecase "Asociar evidencia" as UC4
  usecase "Ejecutar reglas" as UC5
  usecase "Generar alertas" as UC6
  usecase "Revisar caso" as UC7
  usecase "Cerrar alerta" as UC8
  usecase "Generar reportes" as UC9
  usecase "Consultar evidencia" as UC10
  usecase "Configurar reglas" as UC11
  usecase "Auditar actividad" as UC12
}

Fuente --> UC1
Fuente --> UC2
UC2 --> UC3
UC3 --> UC4
UC1 --> UC4
UC4 --> UC5
UC5 --> UC6

Analista --> UC7
Analista --> UC8
Analista --> UC10
Supervisor --> UC7
Supervisor --> UC9
Admin --> UC11
Auditor --> UC9
Auditor --> UC12

@enduml
```

20.8 Diagrama de actividad - Revisión de tasas

```
@startuml
title Actividad - Revisión de tasas de negociación

start

:Ingerir operaciones del día;
:Clasificar por producto, mesa, book y tipo de operación;
:Obtener tasa spot, tasa de comparación y referencias de mercado;
:Calcular spread según compra o venta;
:Excluir operaciones no aplicables;
:Asignar grupo por monto, segmento o franja horaria;
:Calcular percentiles y desviaciones;
:Clasificar operación como no alerta, leve o extremo;

if (Es atípico?) then (Sí)
  :Buscar medio verificable;
  :Validar hora de negociación contra registro;
  :Crear alerta con cálculo y evidencia;
else (No)
  :Guardar resultado analítico;
endif

:Actualizar tablero diario;
stop

@enduml
```

21. Conclusión técnica

La solución viable no es simplemente transcribir llamadas ni crear un chatbot. La solución robusta es una plataforma de cumplimiento con IA que automatice la recolección, normalización, asociación, análisis y trazabilidad de evidencias. El chatbot puede ser útil para consultar casos y hacer preguntas, pero no debe reemplazar el motor de reglas, el repositorio de evidencias ni el flujo de revisión humana.

El primer MVP debe enfocarse en un control con regla clara y alto valor: registro oportuno. Esto permite demostrar rápidamente reducción de trabajo manual, trazabilidad de evidencia, comparación de horarios y generación de alertas explicables. Después se deben incorporar anuladas/modificadas, atribuciones y revisión de tasas, porque esos controles requieren mayor calidad de datos, parametrización y calibración estadística.

La arquitectura debe diseñarse desde el inicio como un sistema auditable. Cada alerta debe poder responder cuatro preguntas: qué operación se evaluó, qué regla se aplicó, qué evidencia la soporta y quién tomó la decisión final. Si el sistema no puede responder eso, no es una solución de compliance; es solo automatización superficial.